Математическая обработка результатов эксперимента

Конспект лекций

Ростов-на-Дону 2021

Математическая обработка результатов эксперимента

1. Среднее арифметическое, мода и медиана. Среднее квадратическое отклонение

Вероятно, Вы отлично знаете, что такое среднее арифметическое. Если мы имеем набор каких-то величин, и все они одной природы (усреднять килограммы с километрами мы, конечно, не можем), надо посчитать сумму, а затем, поделив ее на количество слагаемых, найти среднее арифметическое. Казалось бы, простое и хорошо знакомое действие, но и тут имеется несколько проблем для обсуждения. При знакомстве с некоторыми "показателями" поневоле вспоминается известная шутка о "средней температуре по больнице".

Пример. Допустим, фирма имеет две палатки, торгующие горячей выпечкой, которую они пекут на месте из полуфабрикатов. В таблице приводится примерная сводка ежедневной выручки каждой из палаток за неделю (в руб.).

Дни недели	Понедельник	Вторник	Среда	Четверг	Пятница	Суббота	Воскресенье
Палатка 1	205	268	258	218	341	1515	1397
Палатка 2	759	801	670	599	633	420	301

Различие в ежедневной выручке в основном связано с расположением палаток. Палатка 1 находится в парке отдыха, в то время как Палатка 2 расположена напротив школы и вблизи проходной крупного НИИ.

Владелец фирмы решил выплачивать ежемесячную премию продавцам той палатки, которая даст в этом месяце большую выручку. При распределении премии выяснилась удивительная вещь: выигрыш в этом "соревновании" зависел только от количества выходных в месяце.

Не хотелось бы приводить большое количество цифр за весь месяц в целом, но и без этого видно, что если бы владельцу фирмы пришла в голову идея ежедневного премирования победителя какой-то фиксированной суммой,

"Палатка выходного дня" могла бы рассчитывать на премии в два с половиной раза реже, хотя недельная выручка от нее больше.

В таких условиях более разумное соревнование могло бы быть основано на осреднении показателей за неделю. Допустим, недельные показатели практически совпали. Как оценить, какая из палаток полезнее для фирмы, если по каким-то причинам фирме необходимо продать одну из них?

Если выручка практически совпадает, владелец, по-видимому, поинтересуется стабильностью работы торговой точки. Вины продавцов в этом нет, но если оборудование работает два дня в неделю на износ, а в остальное время больше простоев, выход из строя такого оборудования более вероятен. Пусть в один (случайным образом выпавший) день в неделю идет сильный дождь, и на улицах мало прохожих, падение выручки особенно резко заметно, когда такой дождливый день совпадает с одним из выходных. Для сравнения можно представить спортсменов, которые имеют равные шансы выиграть, но один из них выступает ровнее. Скорее всего, именно он и будет принят в состав сборной.

Но вот еще один вопрос: а не делает ли эта самая нестабильная палатка работу фирмы в целом более стабильной, прекрасно дополняя работу палатки 2? Давайте выдвинем это утверждение в качестве гипотезы и попробуем его доказать или опровергнуть. Чтобы оценить эту проблему количественно, надо прежде всего просуммировать дневную выручку обеих палаток.

Дни недели	Понедельник	Вторник	Среда	Четверг	Пятница	Суббота	Воскресенье
Палатки 1+2	964	1069	928	817	974	1935	1698

То, что мы описали общими словами как "нестабильность работы", в статистике называется **характеристикой рассеивания**. К ним относятся такие показатели как дисперсия и среднее квадратическое отклонение. Покажем на

предыдущем примере, как определяются эти понятия. Посчитаем сначала среднее арифметическое выручки для каждой палатки отдельно, и для обеих палаток вместе (осреднение проводим за семь дней):

$X_{ср.1}$=600 руб., $X_{ср.2}$=598 руб., $X_{ср.1+2}$=1198 руб.

Чтобы сравнить разброс значений, посчитаем для обеих палаток дневные отклонения выручки от их собственного среднего значения.

Чтобы измерить, насколько одна палатка "нестабильнее" другой, хочется сложить всю строку за неделю и получить общее отклонение за весь отчетный период. Но этого делать нельзя, мы сами так построили эти показатели, что, сложив, получим ноль (с точностью до погрешности округления - среднее арифметическое величина не обязательно целая). Чтобы избежать этого обнуления, нам надо, чтобы каждое отклонение от среднего арифметического "лишилось" своего знака. Для этого возводят каждую величину в квадрат, и лишь затем суммируют весь ряд значений.

Дни недели	Понедельник	Вторник	Среда	Четверг	Пятница	Суббота	Воскресенье	ВСЕГО
Палатка 1	-395	-332	-342	-383	-259	915	797	0
Палатка 2	161	203	72	1	35	-178	-297	0
Палатки 1+2	-234	-129	-270	-382	-224	737	500	0

Чтобы не зависеть от периода осреднения делят полученную сумму квадратов на число слагаемых (в нашем случае, по-прежнему на семь). Такая величина называется **дисперсией**.

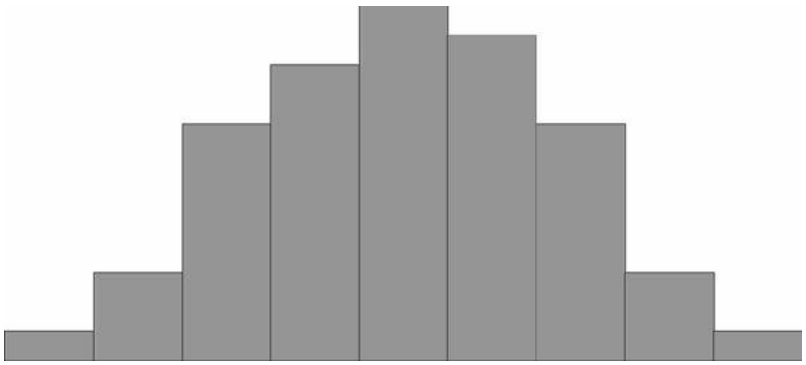
	Дисперсия (руб.^{2})	Среднее квадратическое отклонение (руб.)
Палатка 1	295522	543,6
Палатка 2	27633	166,2
Палатки 1+2	161938	402,4

Мы видим, что дисперсия действительно очень показательная величина. У "Палатки выходного дня" она выше более, чем в десять раз. Дисперсию можно посчитать в Excel автоматически, даже не считая предварительно среднее арифметическое, программа сделает это сама. Для этого, находясь в файле Excel, нажмите в верхнем меню кнопку f_x. Затем, выберите среди функций тип "СТАТИСТИЧЕСКИЕ", и из предложенного перечня в окошке - ДИСПРА. Затем, по подсказке, поставив курсор в поле "Число 1" проведите мышью вдоль строки с набранными значениями. Этот вид подсчета называется "вычисление смещенной дисперсии по генеральной совокупности". Дисперсией часто пользуются, но более удобная характеристика носит название **среднее квадратическое отклонение** (обычно обозначается греческой буквой омега. Среднее квадратическое отклонение - это квадратный корень из дисперсии, он удобен тем, что имеет ту же размерность, что и исходные величины. Так, в нашем случае, дисперсия имела бы размерность "рубли в квадрате", в то время как среднее квадратическое отклонение получается просто и привычно, в рублях. В нашем примере, видно, что суммарная дисперсия и среднее квадратическое отклонение у двух палаток вместе все-таки выше, чем у одной первой палатки, причем среднее квадратическое отклонение выше более, чем в два раза. Значит, наша гипотеза о "повышенной стабильности суммы" за счет присутствия второй палатки несостоятельна. Иногда, вместо среднего арифметического употребляют другие характерные величины, если это по каким-то причинам лучше описывает выборку. Так если расставить выборку по возрастанию (или убыванию) той величины, которой мы интересуемся, то **медиана** - это то, что будет ровно посередине "строя". Например, если мы

расположим по порядку длительности интервалы времени: *секунда, минута, час, сутки и неделя* - то медианой будет час. Еще одно понятие для замены среднего - **мода**. Само название позволяет легко запомнить это определение. Если мы выстроим по порядку все пары обуви на складе по размеру, то самый ходовой размер будет модой. Мода - это то, что непременно должны учитывать производители упаковок и фасовщики. Если бы большинство людей покупало за один раз стакан молока, молочные пакеты не были бы литровыми. В следующем параграфе мы начнем работать со случайными величинами, имеющими нормальное распределение, и эти понятия нам снова встретятся.

2. Нормальное распределение и его свойства

Если выйти на улицу любого города и случайным образом выбранных прохожих спросить о том, какой у них рост, вес, возраст, доход, и т.п., а потом построить график любой из этих величин, например, роста... Но не будем спешить, сначала посмотрим, как можно построить такой график. Сначала, мы просто запишем результаты своего исследования. Потом, мы отсортируем всех людей по группам, так чтобы каждый попал в свой диапазон роста, например, "от 180 до 181 включительно". После этого мы должны посчитать количество людей в каждой подгруппе-диапазоне, это будет частота попадания роста жителей города в данный диапазон. Обычно эту часть удобно оформить в виде таблички. Если затем эти частоты построить по оси y, а диапазоны отложить по оси x, можно получить так называемую **гистограмму**, упорядоченный набор столбиков, ширина которых равна, в данном случае, одному сантиметру, а длина будет равна той частоте, которая соответствует каждому диапазону роста. Если Вам попалось достаточно много жителей, то Ваша схема будет выглядеть примерно так:

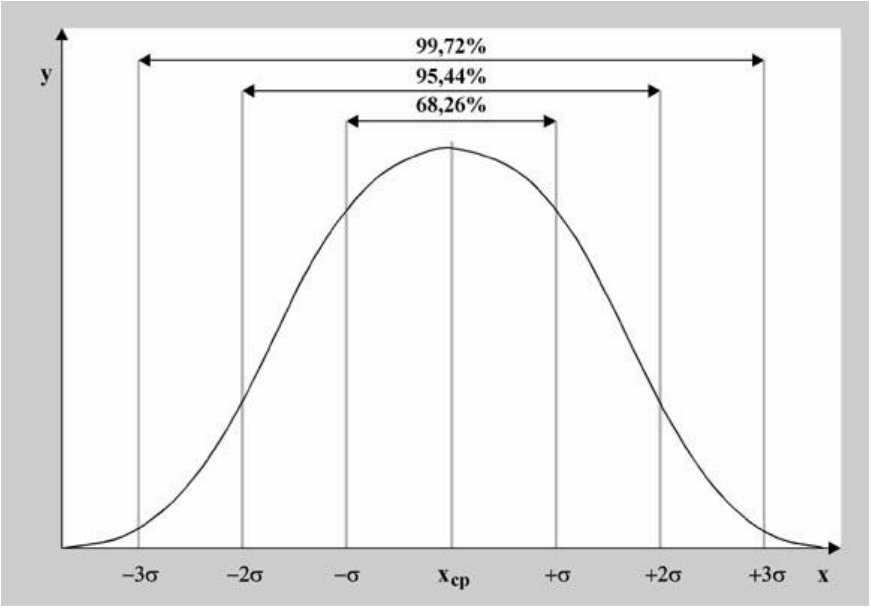


Дальше можно уточнить задачу. Каждый диапазон разбить на десять, жителей рассортировать по росту с точностью до миллиметра. Диаграмма станет глаже, но уменьшится по высоте, "оплывет" вниз, т.к. в каждом маленьком диапазоне количество жителей уменьшается. Чтобы избежать этого, просто увеличим масштаб по вертикальной оси в 10 раз. Если гипотетически повторить эту процедуру несколько раз, будет вырисовываться та знаменитая колоколообразная фигура, которая характерна для нормального (или Гауссова) распределения. В результате, относительная частота встречаемости каждого конкретного диапазона роста может быть посчитана как отношение площади "ломтика" кривой, приходящегося на этот диапазон к площади подо всей кривой. Стандартизарованные кривые нормального распределения, значения функций которых приводятся в таблицах книг по статистике, всегда имеют суммарную площадь под кривой равную единице. Это связано с тем, что, как Вы помните из курса теории вероятности, вероятность достоверного события всегда равна 100% (или единице), а для любого человека иметь хоть какое-то значение роста - достоверное событие. А вот вероятность того, что рост произвольного человека попадет в определенный выбранный нами диапазон, будет зависеть от трех факторов.

Во-первых, от величины такого диапазона - чем точнее наши требования, тем меньше вероятности, что нам повезет.

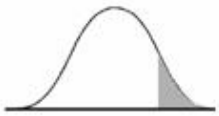
Во-вторых, от того, насколько "популярен" выбранный нами рост. Напомним, что мода - самое часто встречающееся значение роста. Кстати для нормального распределения мода, медиана и среднее значение совпадают. Кривая нормального распределения симметрична относительно среднего

значения. **И, в-третьих,** вероятность попадания роста в определенный диапазон зависит от характеристики рассеивания случайной величины. Отчасти это связано с единицами измерения (представьте, что мы бы измеряли людей в дюймах, а не в миллиметрах, но сами люди и их рост были бы теми же). Но дело не только в этом. Просто некоторые процессы кучнее группируются возле среднего значения, в то время как другие более разбросаны. Например, рост собак и рост домашних кошек имеют разный разброс значений, их кривые нормального распределения будут выглядеть по-разному (напомним еще раз, что площадь под обеими кривыми будет единичной). Так, кривая для роста кошек будет более узкой и высокой, а для роста собак кривая будет ниже и шире. Для характеристики разброса конечного ряда данных в прошлом разделе мы использовали величину среднего квадратического отклонения. Аналогичная величина используется для характеристики кривой нормального распределения. Она обозначается буквой *s* и называется в этом случае **стандартным отклонением.** Это очень важная величина для кривой нормального распределения. Кривая нормального распределения полностью задана, если известно среднее значение $X_{ср.}$ и отклонение *s*. Кроме того, любой житель города с вероятностью 68% попадет в диапазон роста $X_{ср.} \pm s$, с вероятностью 95% - в диапазон $X_{ср.} \pm 2s$, и с вероятностью 99,7% - в диапазон $X_{ср.} \pm 3s$.



Для вычисления других значений вероятности, которые могут Вам понадобиться, можно воспользоваться приведенной таблицей:

Таблица вероятности попадания случайной величины в отмеченный (заштрихованный) диапазон

σ				
0,0	50,00	50,00	100,00	0,00
0,1	46,02	53,98	92,04	7,97
0,2	42,07	57,93	84,14	15,85
0,3	38,21	61,79	76,42	23,58
0,4	34,46	65,54	68,92	31,29
0,5	30,85	69,15	61,70	38,30
0,6	27,43	72,57	54,86	45,15
0,7	24,20	75,80	48,40	51,61
0,8	21,19	78,81	42,38	57,63
0,9	18,41	81,59	36,82	63,19
1,0	15,87	84,13	31,74	68,17
1,1	13,57	86,43	27,14	72,87
1,2	11,51	88,49	23,02	76,99
1,3	9,68	90,32	19,36	80,64
1,4	8,08	91,92	16,16	83,85
1,5	6,68	93,32	13,36	86,64
1,6	5,48	94,52	10,96	89,04
1,7	4,46	95,54	8,92	91,08
1,8	3,59	96,41	7,18	92,81
1,9	2,87	97,13	5,74	94,25
2,0	2,28	97,72	4,56	95,45
2,1	1,79	98,21	3,58	96,43
2,2	1,39	98,61	2,78	97,22
2,3	1,07	98,93	2,14	97,85
2,4	0,82	99,18	1,64	98,36
2,5	0,62	99,38	1,24	98,76
2,6	0,47	99,53	0,94	99,07
2,7	0,35	99,65	0,70	99,31
2,8	0,26	99,74	0,52	99,49
2,9	0,19	99,81	0,38	99,63
3,0	0,14	88,86	0,28	99,73

3. Выборки и доверительные интервалы

Пусть у нас имеется большое количество предметов, с нормальным распределением некоторых характеристик (например, полный склад однотипных овощей, размер и вес которых варьируется). Вы хотите знать средние характеристики всей партии товара, но у Вас нет ни времени, ни желания измерять и взвешивать каждый овощ. Вы понимаете, что в этом нет необходимости. Но сколько штук надо было бы взять на выборочную проверку?

Прежде, чем дать несколько полезных для этой ситуации формул напомним некоторые обозначения.

Во-первых, если бы мы все-таки промерили весь склад овощей (это множество элементов называется генеральной совокупностью), то мы узнали бы со всей доступной нам точностью среднее значение веса всей партии. Назовем это среднее значение $\mathbf{X_{ср.ген.}}$ - генеральным средним. Мы уже знаем, что нормальное распределение определяется полностью, если известно его среднее значение и отклонение s . Правда, пока мы ни $X_{ср.ген.}$, ни sгенеральной совокупности не знаем. Мы можем только взять некоторую выборку, замерить нужные нам значения и посчитать для этой выборки как среднее значение $X_{ср.выб.}$, так и среднее квадратическое отклонение $S_{выб.}$

Известно, что если наша выборочная проверка содержит большое количество элементов (обычно n больше 30), и они взяты **действительно случайным образом**, то s генеральной совокупности почти не будет отличаться от $S_{выб}$

Кроме того, для случая нормального распределения мы можем пользоваться следующими формулами:

С вероятностью 95%

$$X_{ср.ген} = X_{ср.выб.} \pm 1{,}96\frac{\sigma}{\sqrt{n}}$$

С вероятностью 99%

$$X_{ср.ген} = X_{ср.выб.} \pm 2{,}58\frac{\sigma}{\sqrt{n}}$$

В общем виде с вероятностью P(t)

$$X_{ср.ген} = X_{ср.выб.} \pm t\frac{\sigma}{\sqrt{n}} \qquad (1)$$

Связь значения t со значением вероятности P(t), с которой мы хотим знать доверительный интервал, можно взять из следующей таблицы:

P(t)	0.683	0.950	0.954	0.990	0.997
t	1.00	1.96	2.00	2.58	3.00

Таким образом, мы определили, в каком диапазоне находится среднее значение для генеральной совокупности (с данной вероятностью). Если у нас нет достаточно большой выборки, мы не можем утверждать, что генеральная совокупность имеет $s = S_{выб}$. Кроме того, в этом случае проблематична близость выборки к нормальному распределению. В этом случае также пользуются $S_{выб}$ вместо s в формуле:

$$X_{ср.ген} = X_{ср.выб.} \pm t\frac{\sigma}{\sqrt{n}}$$

но значение t для фиксированной вероятности P(t) будет зависеть от количества элементов в выборке n. Чем больше n, тем ближе будет полученный доверительный интервал к значению, даваемому формулой (1). Значения t в этом случае берутся из другой таблицы (t-критерий Стьюдента), которую мы приводим ниже:

Значения t-критерия Стьюдента для вероятности 0,95 и 0,99

n	P		n	P	
	0,95	**0,99**		**0,95**	**0,99**
2	12,71	63,66	18	2,11	2,90
3	4,30	9,93	19	2,10	2,88
4	3,18	5,84	20	2,093	2,861
5	2,78	4,60	25	2,064	2,797
6	2,57	4,03	30	2,045	2,756
7	2,45	3,71	35	2,032	2,720
8	2,37	3,50	40	2,022	2,708
9	2,31	3,36	45	2,016	2,692
10	2,26	3,25	50	2,009	2,679
11	2,23	3,17	60	2,001	2,662
12	2,20	3,11	70	1,996	2,649
13	2,18	3,06	80	1,991	2,640
14	2,16	3,01	90	1,987	2,633
15	2,15	2,98	100	1,984	2,627
16	2,13	2,95	120	1,980	2,617
17	2,12	2,92	∞	1,960	2,576

Пример 3. Из работников фирмы случайным образом отобрано 30 человек. По выборке оказалось, что средняя зарплата (в месяц) составляет 10 тыс. рублей при среднем квадратическом отклонении 3 тыс. рублей. С вероятностью 0,99 определить среднюю зарплату в фирме.

Решение: По условию имеем n = 30, $X_{ср.}$=10000, S=3000, P = 0,99. Для нахождения доверительного интервала воспользуемся формулой, соответствующей критерию Стьюдента. По таблице для n = 30 и P = 0,99 находим t=2,756, следовательно,

$$30000 - 2{,}756 \cdot \frac{5000}{\sqrt{30}} < X_{\text{ср ген}} < 30000 + 2{,}756 \cdot \frac{5000}{\sqrt{30}}$$

т.е. искомый доверительный интервал 27484 < $X_{ср.ген}$ < 32516. Итак, вероятностью 0,99 можно утверждать, что интервал (27484; 32516) содержит внутри себя среднюю зарплату в фирме. Мы надеемся, что Вы будете пользоваться этим методом, при этом не обязательно, чтобы при Вас каждый раз была таблица. Подсчеты можно проводить в Excel автоматически.

Находясь в файле Excel, нажмите в верхнем меню кнопку fx. Затем, выберите среди функций тип "статистические", и из предложенного перечня в окошке - СТЬЮДРАСПОБР. Затем, по подсказке, поставив курсор в поле "вероятность" наберите значение обратной вероятности (т.е. в нашем случае вместо вероятности 0,95 надо набирать вероятность 0,05). Видимо, электронная таблица составлена так, что результат отвечает на вопрос, с какой вероятностью мы можем ошибиться. Аналогично в поле "степень свободы" введите значение (n-1) для своей выборки.

4. Центральная предельная теорема. Систематические изменения или случайность

Мы уже знаем, что нормальное распределение - особенное. Некоторые его свойства мы сможем использовать и для распределений, которые, строго говоря, нормальными не назовешь. Задача, которую мы рассмотрим в этом разделе имеет чрезвычайно важное значение для бизнеса, это задача о диагностировании тенденций к изменению показателей.

Удобство использование нормального распределения некоторых случайных величин и особые возможности, которые закон нормального распределения предоставляет исследователю, породили ряд теорем, которые позволяют пользоваться этими свойствами даже, если генеральная совокупность представляет собой "не вполне нормальное распределение". **Центральная предельная теорема** имеет несколько формулировок, мы не будем их здесь полностью приводить и доказывать. Для нас важно знать только то, что в большинстве случаев среднее арифметическое выборки, взятой из генеральной совокупности (напомним, что это среднее арифметическое - тоже случайная величина), ложится на нормальное распределение гораздо лучше, чем исходная генеральная совокупность. Другими словами, если мы возьмем несколько выборок из генеральной совокупности, то средние арифметические величины этих выборок будут представлять собой новую случайную величину с практически нормальным распределением. Именно эта теорема и позволит нам проверять так называемые статистические гипотезы, т.е. делать заключение о наличии тенденции к изменению показателей деятельности, которые сами по себе, являясь случайными величинами, имеют право на некоторый разброс.

Пример. Фирма поместила информацию о своей продукции в каталоге. Был указан один из двух номеров телефона отдела продаж, на который и раньше поступали звонки потенциальных покупателей. Другой номер телефона в каталоге не упоминался. За два месяца до выхода каталога и в течение двух месяцев после было зарегистрировано следующее количество

звонков на эти телефоны (два столбца в таблице). Как нам определить, подействовала ли информация, данная в каталоге, или мы имеем дело со случайным оживлением на рынке, а деньги на рекламу потрачены напрасно?

	До рекламы	После рекламы	Для сравнения ожидаемая величина (число звонков на телефон из каталога)
Телефон из каталога	216	305	274,5
Другой телефон	142	150	180,5
Всего	**358**	**455**	**(455)**

Последний столбец в таблице - ожидаемые величины. Это наши оценки, сделанные из предположения, что ничего не изменилось, и реклама не оказала никакого действия, т.е. произошло общее оживление на рынке и больше ничего, а пропорции между числом звонков на оба телефона должны сохраниться в точности. {Ожидаемая величина для телефона из каталога}=455г216/358=274,5 {Ожидаемая величина для другого телефона}=455г142/358=180,5. Наше предположение, о том, что реклама не оказала никакого воздействие на изменение числа покупателей, носит название **нулевой гипотезы. Альтернативная гипотеза** заключается в предположении о наличии такого влияния. Наша задача - выбрать более **достоверную** из двух этих гипотез. Чтобы оценить, насколько значимы отклонения реальной ситуации от ожидания по нулевой гипотезе, для обоих телефонов мы должны посчитать величину:

$$\chi^2 = \frac{(\text{Полученное значение} - \text{Ожидаемая величена})^2}{\text{Ожидаемая величена}}$$

поставить их в таблицу и просуммировать.

	До рекламы	После рекламы	Для сравнения ожидаемая величина (телефон из каталога)	χ^2
Телефон из каталога	216	305	274,5	3,38
Другой телефон	142	150	180,5	5,15
Всего	**358**	**455**	**(455)**	**8,53**

Дальнейшие наши действия - определить, с какой вероятностью посчитанные отклонения "ложатся" на соответствующую кривую. Для такой оценки можно воспользоваться значениями так называемого X^2-критерия Пирсона. Обычно эти значения задаются в виде стандартных таблиц в книгах по статистике. Дадим и мы такую таблицу (χ - греческая буква "хи"):

d.f.	P					
	0,99	**0,95**	**0,90**	**0,10**	**0,05**	**0,01**
1	0,0002	0,004	0,02	2,71	3,84	6,64
2	0,02	0,10	0,21	4,61	5,99	9,21
3	0,12	0,35	0,58	6,25	7,82	11,34
4	0,30	0,71	1,06	7,78	9,49	13,28
5	0,55	1,15	1,61	9,24	11,07	15,09
6	0,87	1,64	2,20	10,65	12,59	16,81
7	1,24	2,17	2,83	12,02	14,06	18,48
8	1,65	2,73	3,49	13,36	15,51	20,09
9	2,09	3,33	4,17	14,68	16,92	21,67
10	2,56	3,94	4,87	15,99	18,31	23,21
11	3,05	4,58	5,58	17,28	19,68	24,72
12	3,57	5,23	6,30	18,55	21,03	26,22
13	4,11	5,89	7,04	19,81	22,36	27,68
14	4,66	6,57	7,79	21,06	23,69	29,14
15	5,23	7,26	8,55	22,31	25,00	30,58
16	5,81	7,96	9,31	23,54	26,30	32,00
17	6,41	8,67	10,09	24,77	27,59	33,41
18	7,02	9,39	10,86	25,99	28,87	34,81
19	7,63	10,12	11,65	27,20	30,14	36,19
20	8,26	10,85	12,44	28,41	31,41	37,57
21	8,90	11,59	13,24	29,62	32,67	38,93
22	9,54	12,34	14,04	30,81	33,92	40,29
23	10,20	13,09	14,85	32,01	35,17	41,64
24	10,86	13,85	15,66	33,19	36,42	43,98
25	11,52	14,61	16,47	34,38	37,65	44,31
26	12,20	15,37	17,29	35,56	38,89	45,64
27	12,88	16,15	18,11	36,74	40,11	46,96
28	13,56	16,93	18,94	37,92	41,34	48,28
29	14,26	17,71	19,77	39,09	42,56	49,59
30	14,95	18,49	20,60	40,26	43,77	50,89
40	22,16	26,51	29,05	51,81	55,76	63,69
50	29,71	34,76	37,69	63,17	67,51	76,15
100	70,07	77,93	82,36	118,50	124,34	135,81

Теперь несколько слов о том, как пользоваться этой таблицей. Буквы d.f. **означают число степеней свободы**. Чтобы посчитать степени свободы нужно просто брать в таблице с исходными данными количество строк **n** и столбцов **m**, и посчитать величину **(n-1)·(m-1)**. Это и будет количество степеней свободы в каждом конкретном случае. Правда, строки и столбцы берутся только для самих исходных данных, ни строка суммирования (всего), ни столбец подсчета ожидаемых величин при определении степени свободы не учитывается. В нашем случае **d.f.=(2-1)·(2-1)=1**, это означает, что степень свободы равна единице, и в таблице χ^2 мы должны пользоваться соответствующей строкой (верхней). Теперь о столбцах этой таблицы. Цифры 0,99; 0,95; и т.д. означают, что величины отклонений χ^2, стоящие в этих столбцах с вероятностью 0,99; 0,95; и т.д. возникли случайно. В нашем примере, вероятность случайного происхождения отклонения составляет менее 0,01 (т.е. меньше одного шанса из ста!). Мы вполне можем считать, что реклама оказала воздействие. Обратите внимание, что критерий X^2не говорит категорически, что случайность тут невозможна, просто вероятность этого очень мала. Другими словами, если мы отбросим нулевую гипотезу и выберем альтернативную, то вероятность ошибки будет меньше одного процента.

Если Вы будете пользоваться этим методом, совсем не нужно считать каждый раз вручную все отклонения. Подсчеты можно проводить в Excel автоматически. Сначала запишите известные Вам показатели в виде таблицы. Затем посчитайте в Excel столбец ожидаемых величин. После этого нажмите в верхнем меню кнопку f_x. Затем, выберите среди функций тип "статистические", и из предложенного перечня в окошке - ХИ2ТЕСТ. Затем, по подсказке, поставив курсор в поле "ожидаемый интервал" выделите мышью столбец ожидаемых значений (но не захватывайте сумму в нижней строке). Аналогично в поле "фактический интервал" введите массив из столбика фактических данных после рекламы. Программа сама посчитает граничную вероятность того, что отклонение было случайным. Так в нашем

варианте более точное значение вероятности составляет примерно 0,0035. В таблице мы попали по значению X^2 между столбцами и посчитать вероятность с такой точностью не смогли. Видимо для того, чтобы Вы привыкли пользоваться подобными оценками, имеет смысл обсудить вопрос о "степени свободы". Что это такое и какие степени свободы вообще могут быть? Понятно, что оценка значимости происходящих изменений может происходить только при наличии данных, как полученных при гипотетическом воздействии этих изменений, так и свободных от изменений. В качестве заведомо не подверженных изменениям данных в нашем примере выступали показания числа звонков на оба телефона до публикации каталога. Кроме того, для дополнительной объективности данных, мы использовали один телефон как неизвестный в рекламе. Это позволило нам исключить возможное влияние сезонных изменений спроса или другие подобные факторы. В других ситуациях, мы можем сравнивать динамику спроса на один товар с динамикой спроса на другой, если идет целевая раскрутка этого товара, или же товар входит в моду. И в этой ситуации свойства нормального распределения помогут нам сделать вывод о значимости происходящих изменений. После этого раздела Вам нужно будет выполнить третье письменное задание.

ЗАДАНИЕ

В одной и той же торговой палатке, чередуясь по неделе, работают два разных продавца (А и В). Таблица со значениями недельной выручки (в тыс. руб.) приводится за 8 последних недель.

Неделя	1	2	3	4	5	6	7	8
Продавец А	*119*		*93*		*89*		*94*	
Продавец В		*132*		*91*		*102*		*105*

Ответьте, пожалуйста, на следующие вопросы:

1. Какой продавец работал лучше с точки зрения суммарной выручки?

2. Можно ли утверждать, что это отклонение было случайным?

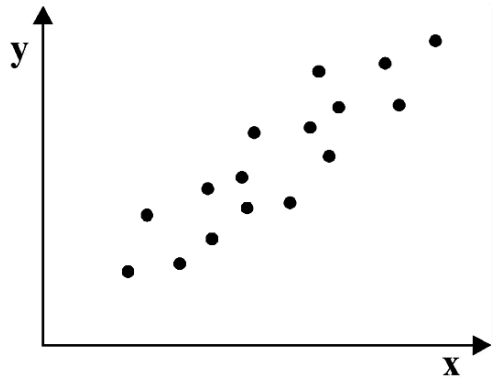
3. С какой достоверностью сделано Ваше утверждение?

5. Введение в корреляционный анализ. Основы регрессионного анализа

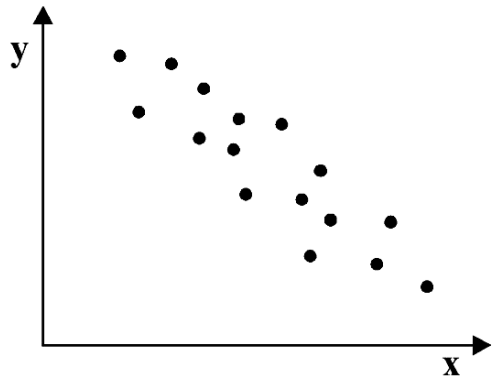
Связь, которая существует между случайными величинами разной природы, например, между величиной X и величиной Y, не обязательно является следствием прямой зависимости одной величины от другой (так называемая функциональная связь). В некоторых случаях обе величины зависят от целой совокупности разных факторов, общих для обеих величин, в результате чего и формируется связанные друг с другом закономерности. Когда связь между случайными величинами обнаружена с помощью статистики, мы не можем утверждать, что обнаружили причину происходящего изменения параметров, скорее мы лишь увидели два взаимосвязанных следствия. Например, дети, которые чаще смотрят по телевизору американские боевики, меньше читают. Дети, которые больше читают, лучше учатся. Не так-то просто решить, где тут причины, а где следствия, но это и не является задачей статистики. Статистика может лишь, выдвинув гипотезу о наличии связи, подкрепить ее цифрами. Если связь действительно имеется, говорят, что между двумя случайными величинами есть корреляция. Если увеличение одной случайной величины связано с увеличением второй случайной величины, корреляция называется прямой. Например, количество прочитанных страниц за год и средний балл (успеваемость). Если, напротив рост одной величины связано с уменьшением другой, говорят об обратной корреляции. Например, количество боевиков и количество прочитанных страниц. Взаимная связь двух случайных величин называется корреляцией, корреляционный анализ позволяет определить наличие такой связи, оценить, насколько тесна и существенна эта связь. Все это выражается количественно. Как определить, есть ли корреляция между величинами? В большинстве случаев, это можно увидеть на обычном графике. Например, по каждому ребенку из нашей выборки можно определить величину X_i (число страниц) и Y_i (средний балл годовой оценки), и записать эти данные в виде таблицы. Построить оси X и Y, а затем нанести на график весь ряд точек таким образом, чтобы каждая из них имела

определенную пару координат (X_i, Y_i) из нашей таблицы. Поскольку мы в данном случае затрудняемся определить, что можно считать причиной, а что следствием, не важно, какая ось будет вертикальной, а какая горизонтальной.

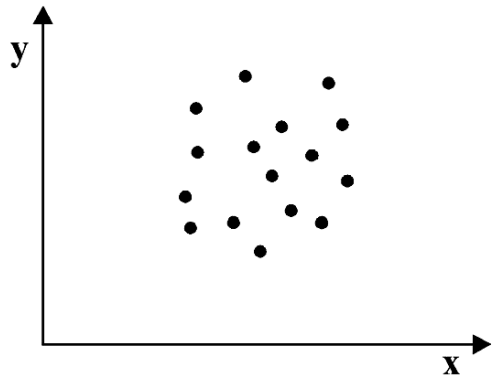
а)



б)



в)



Если график имеет вид а), то это говорит о наличии прямой корреляции, в случае, если он имеет вид б) - корреляция обратная. Отсутствие корреляции тоже можно приблизительно определить по виду графика - это случай в).

С помощью коэффициента корреляции можно посчитать насколько тесная связь существует между величинами.

Пусть, существует корреляция между ценой и спросом на товар. Количество купленных единиц товара в зависимости от цены у разных продавцов показано в таблице:

Цена, руб.	135	140	115	130	120
Число покупок	160	95	200	150	190

Видно, что мы имеем дело с обратной корреляцией. Для количественной оценки тесноты связи используют коэффициент корреляции:

$$r = \frac{\sum_{i=1}^{n}(Xi - Xcp.)(Yi - Ycp.)}{\sqrt{\sum_{i=1}^{n}(Xi - Xcp.)^2 \sum_{i=1}^{n}(Yi - Ycp.)^2}}$$

Коэффициент r мы считаем в Excel, с помощью функции f_x, далее статистические функции, функция КОРРЕЛ. По подсказке программы вводим мышью в два соответствующих поля два разных массива (X и Y). В нашем случае коэффициент корреляции получился r= - 0,988. Надо отметить, что чем ближе к 0 коэффициент корреляции, тем слабее связь между величинами. Наиболее тесная связь при прямой корреляции соответствует коэффициенту r, близкому к +1. В нашем случае, корреляция обратная, но тоже очень тесная, и коэффициент близок к -1. Что можно сказать о случайных величинах, у которых коэффициент имеет промежуточное значение? Например, если бы мы получили r=0,65. В этом случае, статистика позволяет сказать, что две случайные величины частично связаны друг с другом. Скажем на 65% влияние на количество покупок оказывала цена, а на 35% - другие обстоятельства. И еще одно важное обстоятельство надо упомянуть. Поскольку мы говорим о случайных величинах, всегда существует вероятность, что замеченная нами связь - случайное обстоятельство. Причем вероятность найти связь там, где ее нет, особенно

велика тогда, когда точек в выборке мало, а при оценке Вы не построили график, а просто посчитали значение коэффициента корреляции на компьютере. Так, если мы оставим всего две разные точки в любой произвольной выборке, коэффициент корреляции будет равен или +1 или -1. Из школьного курса геометрии мы знаем, что через две точки можно всегда провести прямую линию. Для оценки статистической достоверности факта обнаруженной Вами связи полезно использовать так называемую корреляционную поправку:

$$S_r = \frac{1 - r^2}{\sqrt{n - 1}}$$

Связь нельзя считать случайной, если:

$$\left| \frac{r}{S_r} \right| \geq 3$$

В то время как задача корреляционного анализа - установить, являются ли данные случайные величины взаимосвязанными, цель регрессионного анализа - описать эту связь аналитической зависимостью, т.е. с помощью уравнения. Мы рассмотрим самый несложный случай, когда связь между точками на графике может быть представлена прямой линией. Уравнение этой прямой линии Y=aX+b, где a=Ycp.-bXcp.,

$$b = \frac{\sum_{i=1}^{n} (Xi - Xcp.)(Yi - Ycp.)}{\sum_{i=1}^{n} (Xi - Xcp.)^2}$$

Зная уравнение прямой, мы можем находить значение функции по значению аргумента в тех точках, где значение X известно, а Y - нет. Эти оценки бывают очень нужны, но они должны использоваться осторожно, особенно, если связь между величинами не слишком тесная. Отметим также, что из сопоставления формул для b и r видно, что коэффициент не дает значение наклона прямой, а лишь показывает сам факт наличия связи.

Статистическая обработка экспериментальных данных

1. Предварительная математическая обработка статистических данных

После получения результатов эксперимента для дальнейшего их анализа проводится упорядочение данных, их графическое представление и расчет основных числовых характеристик.

Наблюдаемые значения исследуемого признака X называют вариантами и обозначают $x_1, x_2, \ldots, x_k$, числа их наблюдений называют частотами и обозначают $n_1, n_2, \ldots, n_k$. Общее число наблюдений называют объёмом выборки и обозначают n, $n = n_1 + n_2 + \ldots + n_k$.

Последовательность вариант, записанных в возрастающем порядке, называется вариационным рядом. К характеристикам вариационного ряда относятся:

1) Размах варьирования R — это разность между наибольшим $x_{\max}$ и наименьшим $x_{\min}$ значениями, $R = x_{\max} - x_{\min}$;

2) Мода *Mo* — это варианта, имеющая наибольшую частоту;

3) Медиана *Me* — это варианта, делящая вариационный ряд пополам по числу вариант.

Статистическим распределением выборки называют множество вариант и соответствующих им частот. Обычно статистическое распределение выборки представляют в виде таблицы:

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k

Эмпирической функцией распределения называется числовая функция $F^*(x)$, определяющая относительную частоту события $X < x$. Она вычисляется по формуле:

$$F^*(x)=\frac{n_x}{n}, \tag{1}$$

где n_x — сумма частот вариант, значения которых меньше x, n — объём выборки.

$F^*(x)$ является неубывающей функцией, значения которой принадлежат отрезку $[0,1]$. $F^*(x)$ служит оценкой теоретической функции распределения $F(x)$, определяющей вероятность события $X < x$.

Основными графическими формами представления данных наблюдений являются полигон частот и гистограмма.

Полигоном частот называется ломаная линия, звенья которой соединяют точки с координатами (x_1,n_1), (x_2,n_2), … , (x_k,n_k).

Гистограммой называется ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат интервалы одинаковой длины h, а высотами — плотности интервальных частот n_i/h.

Основными характеристиками выборки являются:

1) **Выборочная средняя** $\overline{x}_в$, вычисляется по формуле:

$$\overline{x}_в=\frac{1}{n}\sum_{i=1}^{k}x_i n_i\,. \tag{2}$$

2) **Выборочная дисперсия** $D_в$**,** вычисляется по формуле:

$$D_в=\frac{1}{n}\sum_{i=1}^{k}x_i n_i^2-\left(\overline{x}_в\right)^2. \tag{3}$$

3) **Исправленная дисперсия** S^2**,** вычисляется по формуле:

$$S^2=\frac{n}{n-1}D_в \tag{4}$$

4) **Выборочное среднее квадратическое отклонение** $\sigma_в$**,** вычисляется по формуле:

$$\sigma_в=\sqrt{D_в}. \tag{5}$$

5) **Исправленное среднее квадратическое отклонение *s***, вычисляется по формуле:

$$s = \sqrt{S^2}. \qquad (6)$$

6) **Коэффициент вариации *V***, вычисляется по формуле:

$$V = \frac{s}{\overline{x}_{\text{в}}} 100\%. \qquad (7)$$

Перечисленные характеристики относятся к точечным оценкам, при малых объёмах выборки предпочтительнее пользоваться интервальными оценками.

Доверительным интервалом для параметра θ, точечной оценкой которого является θ^*, называют интервал $(\theta^* - \delta, \theta^* + \delta)$, содержащий с заданной вероятностью γ значение параметра θ, γ называют надежностью оценки.

Например, в случае нормально распределённой случайной величины доверительный интервал для среднего значения при неизвестном параметре σ определяется формулой:

$$\left(\overline{x}_{\text{в}} - \frac{ts}{\sqrt{n}}, \overline{x}_{\text{в}} + \frac{ts}{\sqrt{n}} \right), \qquad (8)$$

где t — критическая точка распределения Стьюдента с $k = n - 1$ степенями свободы для двусторонней области на уровне значимости $\alpha = 1 - \gamma$ определяется по таблицам, например в [1].

Пример. Статистическая обработка результатов измерений (вычисления выполнять с точностью до двух знаков после запятой)

Даны результаты измерений значений случайной величины *X*.

Составить статистическое распределение выборки и найти:

а) характеристики вариационного ряда: размах варьирования, моду, медиану;

б) эмпирическую функцию распределения и построить ее график;

в) построить полигон частот и гистограмму;

г) выборочную среднюю;

д) выборочную и исправленную дисперсии;

е) выборочное и исправленное средние квадратические отклонения (стандарт);

ж) коэффициент вариации (%);

з) доверительный интервал для среднего значения признака X с надежностью γ =0,95;

12; 9; 16; 17; 10; 9; 15; 12; 15;16; 20; 18; 17; 9; 15; 9; 16; 9; 18; 16

Составим статистическое распределение выборки. Для этого расположим варианты в порядке возрастания:

9; 9; 9; 9; 9; 10; 12; 12; 15; 15; 15; 16; 16; 16; 16; 17; 17; 18; 18; 20

и подсчитаем числа наблюдений каждой варианты — частоты. Получим:

x_i	9	10	12	15	16	17	18	20
n_i	5	1	2	3	4	2	2	1

а) Размах варьирования $R = x_{\max} - x_{\min} = 20 - 9 = 11$; мода *Mo*=9; объём выборки *n*=20, поэтому середина вариационного ряда находится между 10-й и 11-й вариантами в упорядоченном вариационном ряду, и медиана вычисляется как их среднее арифметическое, *Me* = (15+15)/2=15.

б) Эмпирическую функцию распределения найдём по формуле (1):

$x \le 9 \quad F^*(x) = 0;$

$9 < x \le 10 \quad F^*(x) = \frac{5}{20} = 0,25;$

$10 < x \le 12 \quad F^*(x) = \frac{5+1}{20} = 0,3;$

$12 < x \le 15 \quad F^*(x) = \frac{5+1+2}{20} = 0,4;$

$15 < x \le 16 \quad F^*(x) = \frac{5+1+2+3}{20} = 0,55;$

$16 < x \le 17 \quad F^*(x) = \frac{5+1+2+3+4}{20} = 0,75;$

$17 < x \le 18 \qquad F^*(x) = \dfrac{5+1+2+3+4+2}{20} = 0{,}85;$

$18 < x \le 20 \qquad F^*(x) = \dfrac{5+1+2+3+4+2+2}{20} = 0{,}95;$

$x > 20 \qquad F^*(x) = \dfrac{5+1+2+3+4+2+2+1}{20} = 1.$

Построим график (рис. 1)

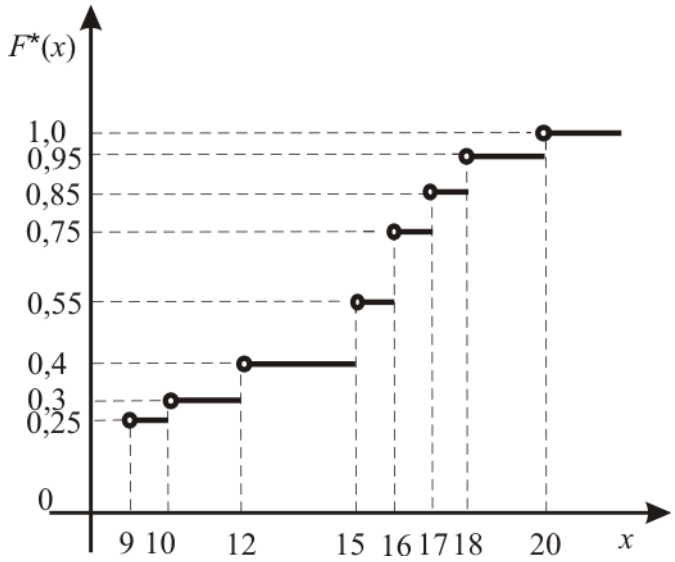


Рис. 1

в) Построим полигон частот (рис. 2). Для этого по оси OX отложим наблюдаемые значения x_i, а по оси OY частоты n_i. Отметим точки с координатами (x_i, y_i) и соединим их последовательно отрезками прямых.

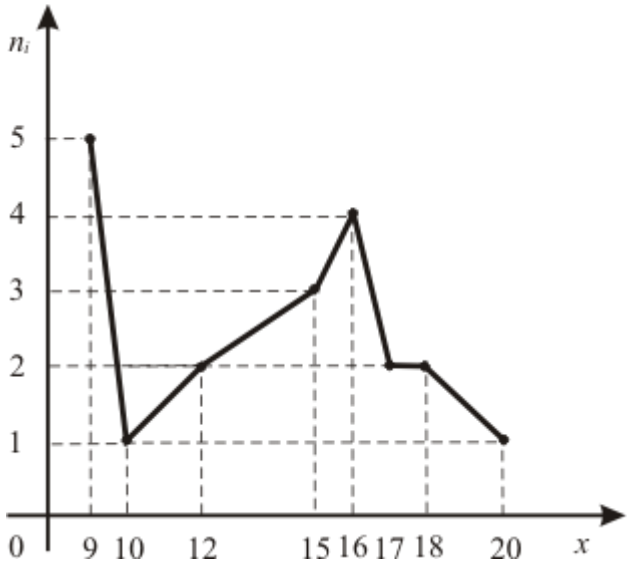


Рис. 2

Для построения гистограммы разобьём интервал изменения x (9,20) на два интервала одинаковой длины h=5,5, подсчитаем интервальные частоты и плотности интервальных частот. Результаты внесём в таблицу 1.

Таблица 1

интервалы	Интервальные частоты n_i	Плотности интервальных частот n_i / h
$[9;14,5)$	8	16/11
$[14,5;\ 20]$	12	24/11

Построим гистограмму (рис. 3).

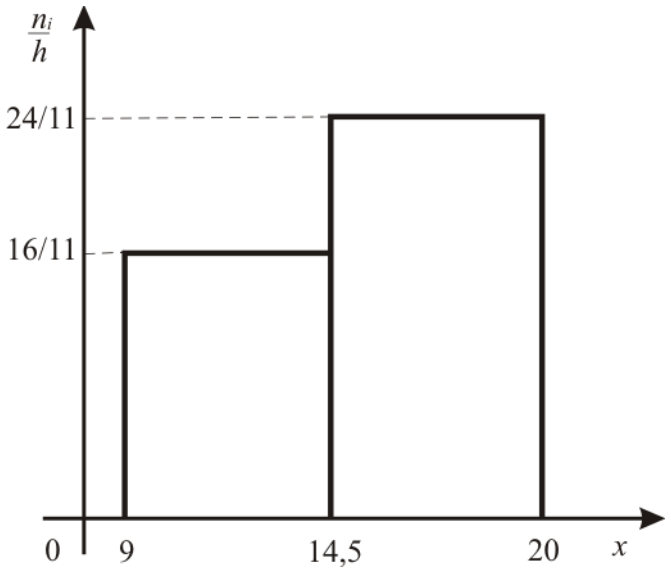


Рис. 3

г) Вычислим выборочную среднюю по формуле (2):

$$\overline{x}_{\text{в}} = \frac{1}{20}\left(9\cdot 5+10\cdot 1+12\cdot 2+15\cdot 3+16\cdot 4+17\cdot 2+18\cdot 2+20\cdot 1\right)=13,9.$$

д) Вычислим выборочную дисперсию формуле (3):

$$D_{\text{в}} = \frac{1}{20}\left(9^2\cdot 5+10^2\cdot 1+12^2\cdot 2+15^2\cdot 3+16^2\cdot 4+17^2\cdot 2+18^2\cdot 2+20^2\cdot 1\right)-$$
$$-13,9^2 = 12,69.$$

Исправленную дисперсию найдём по формуле (4):

$$S^2 = \frac{20}{19}12,69 \approx 13,36.$$

е) Выборочное и исправленное средние квадратические отклонения найдём по формулам (5) и (6):

$\sigma_в = \sqrt{12,69} \approx 3,56; \quad s = \sqrt{13,36} \approx 3,66.$

ж) Коэффициент вариации вычислим по формуле (7):

$$V = \frac{3,66}{13,9} \cdot 100\% = 26,33\%.$$

з) Доверительный интервал для среднего значения признака X найдём по формуле (8). Сначала по таблице [1] найдём критическую точку распределения Стьюдента с числом степеней свободы $k = n - 1 = 20 - 1 = 19$ и уровнем значимости $\alpha = 1 - \gamma = 1 - 0,95 = 0,05.$ Получим t = 2,09 и подставим в формулу (8):

$\left(13,9 - \frac{2,09 \cdot 3,66}{\sqrt{20}}; \quad 13,9 + \frac{2,09 \cdot 3,66}{\sqrt{20}}\right)$. После вычисления получим доверительный интервал для среднего значения $(12,19; \quad 15,61)$.

2. Вычисление ошибок прямых измерений

Ошибки измерений классифицируют как систематические, случайные и грубые промахи.

Систематическими называют такие ошибки, которые возникают из-за известных причин, действующих по определённым законам и, как правило, в определённом направлении. Их можно количественно определить и вносить в измерения соответствующие поправки.

Случайными называют такие ошибки, причины которых неизвестны и которые невозможно учесть заранее. Такие ошибки можно выразить несколькими способами. Часто пользуются понятием предельной ошибки $\Delta_п$, под которой понимают наибольшую случайную ошибку при пользовании исправным прибором при устранённых систематических ошибках. Она может быть определена из паспорта прибора или принята равной половине наименьшего деления шкалы прибора.

При определении величины случайных ошибок можно пользоваться статистической ошибкой, полученной неоднократными измерениями обработкой результатов методами математической статистики. В этом случае последовательность определения случайных ошибок следующая:

1) Прибором измеряют несколько раз (n раз) практически постоянную величину и находят её среднее арифметическое:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \qquad (9)$$

2) Вычисляют исправленную дисперсию измеряемой величины:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \qquad (10)$$

и исправленное среднее квадратическое отклонение (стандарт):

$$s = \sqrt{S^2}. \qquad (11)$$

3) Тогда наибольшая возможная статистическая ошибка с вероятностью 99,73% в случае нормального закона распределения случайной величины будет:

$$\Delta_{\text{п}} = \pm 3s, \qquad (12)$$

а относительная ошибка:

$$\frac{\Delta_{\text{п}}}{\overline{x}} \cdot 100\% = \pm\frac{3s}{\overline{x}} \cdot 100\%. \qquad (13)$$

Пример. Определение погрешности прямых измерений (вычисления выполнять с точностью до двух знаков после запятой)

Даны результаты 10 равноточных измерений некоторой физической величины, проведенные без систематических ошибок. Вычислить

1) среднее значение измеряемой величины;

2) среднеквадратическую ошибку;

3) предельную относительную вероятностную ошибку, предполагая, что результаты измерений распределены нормально;

4) доверительный интервал для истинного значения измеряемой величины с надежностью $\gamma = 0{,}9$.

Результаты измерений:

7,94 8,45 9,09 8,71 8,39 9,37 9,26 8,68 8,28 8,39

1) Найдём среднее арифметическое по формуле (9):

$$\overline{x} = \frac{1}{10}(7,94 + 8,45 + 9,09 + 8,71 + 8,39 + 9,37 + 9,26 + 8,68 + 8,28$$
$$+8,39) = 8,656.$$

2) Вычислим исправленную дисперсию по формуле (10) и исправленное среднее квадратическое отклонение по формуле (11):

$$S^2 = \frac{1}{9}((7,94 - 8,656)^2 + (8,45 - 8,656)^2 + (9,09 - 8,656)^2 +$$
$$+(8,71 - 8,656)^2 + (8,39 - 8,656)^2 + (9,37 - 8,656)^2 + (9,26 - 8,656)^2 +$$
$$+(8,68 - 8,656)^2 + (8,28 - 8,656)^2 + (8,39 - 8,656)^2) = 0,212;$$

$s = \sqrt{0,212} = 0,46$. Итак, среднеквадратическая ошибка равна 0,46

3) Вычислим предельную ошибку по формуле (12) и относительную ошибку по формуле (13):

$$\Delta_{\text{п}} = \pm 3 \cdot 0,46 = \pm 1,38;$$

$$\frac{\Delta_{\text{п}}}{\overline{x}} \cdot 100\% = \pm \frac{1,38}{8,656} \cdot 100\% = \pm 15,94\%.$$

Окончательно результат измерений представляем в виде:

$x = 8,656 \pm 1,38;$ относительная ошибка составляет $\pm 15,94\%$.

4) Доверительный интервал для среднего значения измеряемой величины найдём по формуле (8). Сначала по таблице [1] найдём критическую точку распределения Стьюдента с числом степеней свободы $k = n - 1 = 10 - 1 = 9$ и уровнем значимости $\alpha = 1 - \gamma = 1 - 0,9 = 0,1$. Получим t = 1,83 и подставим в формулу (8):

$\left(8,656 - \frac{1,83 \cdot 0,46}{\sqrt{10}};\quad 8,656 + \frac{1,83 \cdot 0,46}{\sqrt{10}}\right)$. После вычисления получим доверительный интервал для среднего значения $(8,39;\quad 8,92)$.

3. Вычисление ошибок косвенных измерений

В большинстве случаев в ходе эксперимента несколькими приборами измеряются несколько величин и для получения конечного результата эти измерения необходимо обработать, используя математические операции: сложения, умножения и т.д. Поэтому необходимо оценивать точность опыта в целом с помощью вычисления предельной и среднеквадратической ошибок опыта.

Правила вычисления предельной относительной ошибки опыта:

1. Ошибка суммы заключена между наибольшей и наименьшей из относительных ошибок слагаемых. Обычно учитывается или наибольшая ошибка или средняя арифметическая величина (в лабораторной работе будем пользоваться средней арифметической величиной).
2. Ошибка произведения или частного равна сумме относительных ошибок сомножителей или соответственно делимого и делителя.
3. Ошибка *n*-ой степени основания в *n* раз больше относительной ошибки основания.

Для вычисления среднеквадратической ошибки результата косвенных измерений необходимо обеспечить независимость результатов измерений. В этом случае среднеквадратическая ошибка вычисления величины W, являющейся функцией измеряемых прямо параметров $x, y, z, \ldots$ $W = f(x, y, z, ...)$, определяется формулой:

$$S_w = \sqrt{(f'_x)_0^2 S_x^2 + (f'_y)_0^2 S_y^2 + (f'_z)_0^2 S_z^2 + ...}, \qquad (14)$$

где $(f'_x)_0, (f'_y)_0, (f'_z)_0, ...$ — частные производные функции $W = f(x, y, z, ...)$, вычисленные при средних значениях параметров $x, y, z, \ldots$, $S_x^2, S_y^2, S_z^2, ...$ — исправленные дисперсии соответственно $x, y, z, \ldots$.

Пример. Определение погрешности косвенных измерений

В результате многократных измерений были получены средние значения и среднеквадратические ошибки 3-х взаимно независимых

параметров: $\bar{x} = 12,52$; $\bar{y} = 4,58$; $\bar{z} = 7,12$; $s_x = \pm 0,05$; $s_y = \pm 0,015$; $s_z = \pm 0,02$.

Найти:

а) предельную относительную ошибку измерений x, y, z и предельную относительную ошибку определения функции $w = f(x, y, z)$;

б) среднее значение и среднеквадратическую ошибку определения функции $w = f(x, y, z)$.

$\bar{x} = 12,52$; $\bar{y} = 4,58$; $\bar{z} = 7,12$; $s_x = \pm 0,05$; $s_y = \pm 0,015$; $s_z = \pm 0,02$.

$$W = \frac{x - y^3}{z}$$

а) Найдём предельные относительные ошибки измерений x, y, z по формуле (13):

$$\frac{\Delta_п}{\bar{x}} \cdot 100\% = \pm \frac{3s_x}{\bar{x}} \cdot 100\% = \pm \frac{3 \cdot 0,05}{12,52} \cdot 100\% = 1,2\%;$$

$$\frac{\Delta_п}{\bar{y}} \cdot 100\% = \pm \frac{3s_y}{\bar{y}} \cdot 100\% = \pm \frac{3 \cdot 0,015}{4,58} \cdot 100\% = 0,98\%;$$

$$\frac{\Delta_п}{\bar{z}} \cdot 100\% = \pm \frac{3s_z}{\bar{z}} \cdot 100\% = \pm \frac{3 \cdot 0,02}{7,12} \cdot 100\% = 0,84\%.$$

Предельную относительную ошибку определения функции $W = f(x, y, z, ...)$, найдём по правилам вычисления предельной относительной ошибки опыта:

$$\frac{\Delta_п}{\bar{w}} \cdot 100\% = \pm \left(\frac{1}{2} \left(\frac{\Delta_п}{\bar{x}} \cdot 100\% + 3 \cdot \frac{\Delta_п}{\bar{y}} \cdot 100\% \right) + \frac{\Delta_п}{\bar{z}} \cdot 100\% \right) =$$

$$= \left((1,2\% + 3 \cdot 0,98\%) / 2 + 0,84\% \right) = 2,91\%.$$

б) Вычислим среднее значение функции $W = \frac{x - y^3}{z}$:

$$\overline{W} = \frac{x - \bar{y}^3}{\bar{z}} = \frac{12,52 - 4,58^3}{7,12} = -11,73.$$

Для вычисления среднеквадратической ошибки определения функции $W = f(x, y, z, ...)$, по формуле (14) найдём частные производные:

$$f'_x = \frac{1}{z}, \quad f'_y = -\frac{3y^2}{z}, \quad f'_z = -\frac{x - y^3}{z^2}$$

и вычислим их при средних значениях *x*, *y*, *z*:

$$(f'_x)_0 = \frac{1}{7,12} = 0,14; \quad (f'_y)_0 = -\frac{3 \cdot 4,58^2}{7,12} = -8,84;$$

$$(f'_z)_0 = -\frac{12,52 - 4,58^3}{7,12^2} = 1,65.$$

Подставляя в формулу (14), получим:

$$S_w = \sqrt{0,14^2 \cdot 0,05^2 + (-8,84)^2 \cdot 0,015^2 + 1,65^2 \cdot 0,02^2} = 0,137.$$

4. *Расчёт характеристик линейной регрессионной модели*

Одним из эффективных методов установления взаимосвязей между факторами является корреляционно-регрессионный анализ.

Задача корреляционно-регрессионного метода заключается в нахождении эмпирического уравнения, характеризующего связь результативного параметра *Y* с определённым входным фактором *X*.

В качестве формы связи *Y* и *X* широко используют линейную зависимость в силу её простоты в расчётах, а также в связи с тем, что к ней можно привести многие другие виды зависимости.

Расчёт линейной регрессионной модели включает следующие этапы:

1. Расчёт теоретического уравнения линейной регрессии;
2. Оценка силы связи, расчёт коэффициента корреляции;
3. Оценка значимости коэффициента корреляции;
4. Оценка значимости коэффициентов уравнения регрессии;
5. Определение адекватности уравнения регрессии и доверительных границ.

Линейная регрессия *Y* на *X* имеет вид:

$$y(x) = \alpha + \beta x,$$

где α и β — параметры регрессии (β называется коэффициентом регрессии).

Статистические оценки α^* и β^* параметров регрессии α и β выбираются таким образом, чтобы значения $y_i\ (i = 1,2,...,n)$, вычисленные по формуле $y_i = \alpha^* + \beta^* x_i$, были как можно ближе к эмпирическим значениям y_i. В качестве меры близости выбирают сумму квадратов отклонений $\sum_{i=1}^{n}\left(y_i - y_i\right)^2$. Метод нахождения параметров с помощью минимизации суммы квадратов отклонений эмпирических значений y_i от теоретических значений y_i в тех же точках называют методом наименьших квадратов.

Оптимальные значения параметров, полученные согласно этому методу, определяются формулами:

$$\alpha^* = \overline{y} - \beta^* \overline{x}, \quad \beta^* = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}, \tag{15}$$

где $\overline{x}$ и $\overline{y}$ — средние значения *X* и *Y*, которые вычисляют по формулам:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{16}$$

Учитывая (15), запишем эмпирическую линию регрессии в виде:

$$y = \overline{y} + \beta^*\left(x - \overline{x}\right). \tag{17}$$

Силу линейной корреляционной зависимости *Y* и *X* характеризует коэффициент корреляции *r*. Коэффициент *r* изменяется в пределах от -1 до 1. Чем ближе он к ± 1, тем сильнее линейная связь *Y* и *X*, в предельном случае, если $r = \pm 1$, имеет место точная линейная функциональная зависимость *Y* от *X*. Если $r = 0$, то *Y* и *X* не коррелируют. Оценкой коэффициента корреляции *r* служит **выборочный коэффициент корреляции** r^*, который вычисляется по формуле:

$$r^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \qquad (18)$$

Коэффициент корреляции r^*, определяемый по выборочным данным, может не совпадать с действительным значением, соответствующим генеральной совокупности. Для проверки статистической гипотезы о значимости выборочного коэффициента корреляции используют *t*-критерий Стьюдента, наблюдаемое значение которого вычисляется по формуле:

$$t_{\text{набл}} = \frac{r^* \sqrt{n-2}}{\sqrt{1-r^{*2}}}. \qquad (19)$$

Критическое значение *t*-критерия $t_{\text{кр}}$ для числа степеней свободы $k = n - 2$ и уровня значимости α находят по таблицам критических точек распределения Стьюдента [1]. Если $|t_{\text{набл}}| > |t_{\text{кр}}|$, то предположение о нулевом значении коэффициента корреляции не подтверждается, и выборочный коэффициент корреляции значим. Если $|t_{\text{набл}}| < |t_{\text{кр}}|$, то величина r близка к нулю.

Для оценки параметров, входящих в уравнение регрессии (16) , при решении практических задач можно ограничиться построением доверительных интервалов. Для заданной надёжности γ доверительные интервалы для параметров $\bar{y}$ и β определяются формулами:

$$\left(\bar{y} - \frac{t_{\text{кр}} s_{\text{ост}}}{\sqrt{n}};\quad \bar{y} + \frac{t_{\text{кр}} s_{\text{ост}}}{\sqrt{n}}\right), \qquad (20)$$

$$\left(\beta^* - \frac{t_{\text{кр}} s_{\text{ост}}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}};\quad \beta^* + \frac{t_{\text{кр}} s_{\text{ост}}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right), \qquad (21)$$

где $t_{\text{кр}}$ — критическое значение *t*-критерия для числа степеней свободы $k = n - 2$ и уровня значимости $\alpha = 1 - \gamma$, которое находят по таблицам

критических точек распределения Стьюдента [1], $s_{\text{ост}}$ — квадратный корень из остаточной дисперсии $S^2_{\text{ост}}$, которая находится по формуле:

$$S^2_{\text{ост}} = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - y_i\right)^2. \qquad (22)$$

После получения эмпирического уравнения регрессии, проверяют насколько оно соответствует результатам наблюдений. Для проверки гипотезы о значимости уравнения регрессии используют *F*-критерий Фишера, наблюдаемое значение которого вычисляют по формуле:

$$F_{\text{набл}} = \frac{S^2_y}{S^2_{\text{ост}}}, \qquad (23)$$

где S^2_y — исправленная дисперсия *Y*, которая вычисляется по формуле:

$$S^2_y = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2. \qquad (24)$$

Критическое значение *F*-критерия $F_{\text{кр}}$ для числа степеней свободы $k_1 = n-1$ и $k_2 = n-2$ и уровня значимости α находят по таблицам критических точек распределения Фишера-Снедекора [1]. Если $F_{\text{набл}} > F_{\text{кр}}$, то гипотеза о незначимости уравнения регрессии не подтверждается, и уравнение соответствует результатам наблюдений. Если $F_{\text{набл}} < F_{\text{кр}}$, то полученное уравнение незначимо.

Ещё одной характеристикой меры того, насколько эмпирическое уравнение хорошо описывает данную систему наблюдений, является коэффициент детерминации *d*, который вычисляется по формуле:

$$d = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}. \qquad (25)$$

Чем ближе коэффициент *d* к единице, тем лучше описание.

После того как модель построена, она используется для анализа и прогноза. Прогноз осуществляется подстановкой фактора $x = x_0$ в уравнение (17). Получается точечная оценка y:

$$y = \bar{y} + \beta^*(x_0 - \bar{x}). \qquad (26)$$

Доверительный интервал для прогнозируемого значения имеет вид:

$$\left(y - t_{\text{кр}} s_{\text{ост}} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}};\quad y + t_{\text{кр}} s_{\text{ост}} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \right), \qquad (27)$$

где $t_{\text{кр}}$ — критическое значение t-критерия для числа степеней свободы $k = n - 2$ и уровня значимости $\alpha = 1 - \gamma$, которое находят по таблицам критических точек распределения Стьюдента [1].

Пример. Построение модели линейной регрессии

По данным наблюдений определить параметры линейного уравнения регрессии *Y* на *X*. Найти коэффициенты регрессии и корреляции проверить гипотезу о значимости выборочного коэффициента корреляции. Найти доверительные интервалы для параметров уравнения регрессии. Определить коэффициент детерминации. Проверить гипотезу о значимости полученного уравнения регрессии. Найти прогнозируемое моделью значение *y* при $x=x_0$ и найти для него доверительный интервал. Уровень значимости α принять равным 0,05.

X	73	85	102	115	122	126	134	147
Y	0,5	0,7	0,9	1,1	1,4	1,4	1,7	1,9

$x_0 = 140$

Для получения параметров уравнения регрессии составим таблицу.

Таблица 2

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$	y	$(y-y)^2$

73	0,5	-40	-0,7	1600	0,49	28	0,43	0,0049
85	0,7	-28	-0,5	784	0,25	14	0,661	0,0015
102	0,9	-11	-0,3	121	0,09	3,3	0,998	0,0077
115	1,1	2	-0,1	4	0,01	-0,2	1,239	0,0193
122	1,4	9	0,2	81	0,04	1,8	1,373	0,0007
126	1,4	13	0,2	169	0,04	2,6	1,450	0,0025
134	1,7	21	0,5	441	0,25	10,5	1,604	0,0092
147	1,9	34	0,7	1156	0,49	23,8	1,854	0,0021
904	9,6	0	0	4356	1,66	83,8		0,0479

В последней строке таблицы приведены суммы столбцов, используемых в расчётах.

Найдём средние значения X и Y по формуле (16):

$$\overline{x} = \frac{904}{8} = 113; \quad \overline{y} = \frac{9,6}{8} = 1,2.$$

Вычислим коэффициент регрессии по формуле (15):

$\beta^* = \dfrac{83,8}{4356} = 0,01924$ и получим эмпирическое уравнение регрессии, подставляя $\overline{x}, \overline{y}, \beta^*$ в (17):

$$y = 1,2 + 0,01924(x - 113). \qquad (28)$$

По формуле (28) вычислим теоретические значения y и заполним два последних столбца таблицы 2.

Вычислим коэффициент корреляции по формуле (18):

$r^* = \dfrac{83,8}{\sqrt{4356 \cdot 1,66}} = 0,985$ и проверим гипотезу о его значимости. Наблюдаемое значение критерия найдём по формуле (19):

$t_{\text{набл}} = \dfrac{0,985\sqrt{8-2}}{\sqrt{1-0,985^2}} = 13,98.$ По таблице критических точек распределения Стьюдента [1] найдём критическую точку распределения Стьюдента с числом степеней свободы $k = n - 2 = 8 - 2 = 6$ и уровнем значимости $\alpha = 0,05.$

Получим $t_{кр} = 2,45$ и сравним $t_{набл}$ и $t_{кр}$: $|t_{набл}| > t_{кр}$, следовательно, коэффициент корреляции значим, и *Y* и *X* связаны линейной корреляционной зависимостью.

Для определения доверительных интервалов параметров уравнения линейной регрессии (28) найдём остаточную дисперсию по формуле (22):

$S^2_{ост} = \frac{1}{8-2} 0,0479 = 0,008$. Подставляя в формулу (20), получим доверительный интервал для $\overline{y}$: $\left(1,2 - \frac{2,45\sqrt{0,008}}{\sqrt{8}};\ \ 1,2 + \frac{2,45\sqrt{0,008}}{\sqrt{8}}\right)$.

Вычисляя, получим интервальную оценку для $\overline{y}$ с надёжностью $\gamma = 1 - \alpha = 0,95$: $1,123 < \overline{y} < 1,277$.

Доверительный интервал для β получим по формуле (21):

$\left(0,01924 - \frac{2,45\sqrt{0,008}}{\sqrt{4356}};\ \ 0,01924 + \frac{2,45\sqrt{0,008}}{\sqrt{4356}}\right)$. Итак, интервальная оценка для параметра β с надёжностью $\gamma = 0,95$: $0,0159 < \beta^* < 0,0226$.

Проверим гипотезу о значимости полученного уравнения регрессии. Для вычисления наблюдаемого значения *F*-критерия найдём исправленную дисперсию *Y* по формуле (24): $S^2_y = \frac{1}{8-1} \cdot 1,66 = 0,237$. Подставляя в формулу (23), получим: $F_{набл} = \frac{0,237}{0,008} = 29,625$. По таблице критических точек распределения Фишера-Снедекора [1] для числа степеней свободы $k_1 = 8 - 1 = 7$ и $k_2 = 8 - 2 = 6$ на уровне значимости $\alpha = 0,05$ найдём $F_{кр} = 4,21$. Сравнивая наблюдаемое и критическое значения *F*-критерия, получим $F_{набл} > F_{кр}$, следовательно, уравнение значимо.

Для оценки адекватности линейной модели наблюдаемым значениям найдём также коэффициент детерминации по формуле (25):

$d = 1 - \frac{0,0479}{1,66} = 0,971.$ Этот результат истолковывается так: 97,1% изменчивости Y объясняется изменением фактора X, а на остальные случайные факторы приходится 2,9% изменчивости. Однако, этот вывод действителен только для рассматриваемого интервала значений X.

Используем уравнение (28) для прогноза. При $x = x_0 = 140$ точечную оценку для y получим путём подстановки $x = 140$ в формулу (28):

$y = 1,2 + 0,01924(140 - 113) = 1,72.$ Доверительный интервал для y получим по формуле (27):

$$\left(1,72 - 2,45\sqrt{0,008}\sqrt{\frac{1}{8} + \frac{(140-113)^2}{4356}};\quad 1,72 + 2,45\sqrt{0,008}\sqrt{\frac{1}{8} + \frac{(140-113)^2}{4356}}\right).$$

Окончательно, интервальная оценка для y с надёжностью $\gamma = 0,95$:

$1,625 < y < 1,815.$